# A Primal-Dual Method for Large-Scale Image Reconstruction in Emission Tomography *

Calvin A. Johnson† and Ariela Sofer‡

November 13, 1997

**Abstract**

In emission tomography, images can be reconstructed from a set of measured projections using a maximum likelihood (ML) criterion. In this paper, we present a primal-dual algorithm for large-scale three-dimensional image reconstruction. The primal-dual method is specialized to the ML reconstruction problem. The reconstruction problem is extremely large; in several of our data sets the Hessian of the objective function is the product of a 1.4 million by 63 million matrix and its scaled transpose. As such, we only consider approaches that are suitable for large-scale parallel computation. We apply a stabilization technique to the system of equations for computing the primal direction and demonstrate the need for stabilization when approximately solving the system using an early-terminated conjugate gradient iteration.

We demonstrate that the primal-dual method for this problem converges faster than the logarithmic barrier method and considerably faster than the expectation maximization algorithm. The use of extrapolation in conjunction with the primal-dual method further reduces the overall computation required to achieve convergence.

**Key words.** tomography, estimation, large-scale problems, parallel computation, applications of nonlinear programming, primal-dual methods

**AMS subject classifications.** 92C55, 62G05, 65Y05, 90C06, 90C90, 90C30, 49M05

## 1 Introduction

In this paper we consider the image reconstruction problem in emission tomography. This problem is encountered in the field of nuclear medicine, which is concerned with the study of organ function through radioactively labeled "tracer" compounds. The quantity of interest in this problem is the spatial concentration of radioactive emissions within the object under study. The quality of the reconstructed image can depend upon a number of factors including the number of emission events (i.e., counts) collected by the scanner and the method used to reconstruct the image. In studies that are characterized by poor counting statistics (that is, few counts), statistical reconstruction methods that model the Poisson nature of the emission process have been shown to improve image quality over traditional, non-statistical reconstruction methods [31, 51]. The low-count problem has generated considerable interest in the medical imaging community because low radiotracer doses and short scanning durations are highly desirable.

*Ariela Sofer is partly supported by National Science Foundation grant DMI-9414355.

†Division of Computer Research and Technology, National Institutes of Health, Bethesda, Maryland, 20892-5624 (johnson@helix.nih.gov).

‡Department of Operations Research and Engineering, George Mason University, Fairfax, Virginia 22030-4444 (asofer@gmu.edu).

The estimation of emission density in an organ is an inherently three-dimensional (3-D) process. Volume, or 3-D acquisition improves the counting statistics compared with 2-D acquisition (in which axially oblique coincidences are either physically or electronically blocked from detection) but increases the problem size considerably. Since the 3-D problem may involve image and measurement vectors with millions of elements, the amount of computation required to perform 3-D statistical reconstructions can be quite substantial. In our computational studies for example, the larger reconstructions consist of 1.4 million image variables which are reconstructed from a measurement vector with 63 million elements. As such, it is important to use reconstruction methods that converge rapidly. The statistical image reconstruction problem can be posed as a constrained nonlinear optimization problem. In this paper we present a primal-dual method for performing statistical 3-D reconstructions in emission tomography that has been specialized to the intricacies of the application. We demonstrate the rapid convergence of our primal-dual method in computational studies on low-count, 3-D positron emission tomography (PET) data.

This paper is organized as follows. In Section 2 we present the statistical model and develop the objective function. Section 3 reviews the EM method for ML reconstruction. In Section 4 we develop a primal-dual method for ML reconstruction and discuss initialization, stabilization, and extrapolation enhancements. Computational tests comparing the primal-dual results to a logarithmic barrier approach and the EM method on small animal data are presented in Section 5. Some concluding remarks are made in Section 6.

## 2 Statistical model and objective function

We begin our discussion by forming a finite parameter space for the image estimates, as is customary [19]. Consider the situation depicted in Figure 1 where a grid of boxes or *voxels* has been imposed over the emitting object (for simplicity, the Figure is depicted in 2-D; the concept is readily extended to 3-D). Given a set of measurements along lines of coincidence, we seek to estimate $x_i = E\{\xi_i\}\ i = 1, \ldots n$, the expected number of counts emitted from voxel $i$. The number of radioactive events emitted from voxel $i$ is assumed to be a Poisson-distributed random variable with mean $x_i$ [47]. A *system matrix* $C \in \Re^{n \times N}$ is used to model a number of physical effects including spatially dependent resolution and attenuation. The elements $C_{i,j}$ of the system matrix represent the probability that an event emitted from voxel $i$ will be detected by detector pair (coincidence line) $j$. By a Bernoulli thinning process with probability $C_{i,j}$, the number of events emitted from voxel $i$ and detected at coincidence line $j$, $\mathbf{\Xi}_{i,j} = \xi_i C_{i,j}$ is also an independent Poisson variable. The measurements $y_j$ are thus realizations of sums of independent Poisson variables $\mathbf{y}_j = \sum_i \mathbf{\Xi}_{i,j}$ with means $\hat{y}_j = E\{\mathbf{y}_j\} = \sum_i C_{i,j} x_i$. The above is a considerably simplified model of the actual measurement process; for further discussion on its validity to the present situation, see [23].

Given our simplified Poisson model, the likelihood may be written as

$$P\{y|x\} = \prod_j \frac{e^{-\hat{y}_j} \hat{y}_j^{y_j}}{y_j!} = \prod_j \frac{e^{-\sum_i C_{i,j} x_i} \left(\sum_i C_{i,j} x_i\right)^{y_j}}{y_j!}.$$

The ML objective function is formed by taking the log likelihood

$$f_{ML}(x) = \log P\{y|x\} = \sum_j \left( -\sum_i C_{i,j} x_i + y_j \log \sum_i C_{i,j} x_i - \log(y_j)! \right),$$

or (ignoring the constant term),

$$f_{ML}(x) = \sum_j \left( -\left(C^T x\right)_j + y_j \log \left(C^T x\right)_j \right) = -q^T x + \sum_j y_j \log \left(C^T x\right)_j, \tag{1}$$

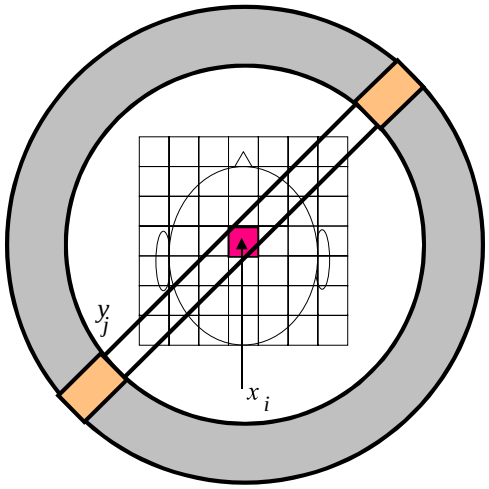


Figure 1: Relationship between estimate $x_i$ and measurement $y_j$. Shown here is the case of PET, where emission-count measurements are taken along coincidence lines from pairs of detectors. A finite parameter space is formed by imposing a grid of voxels over the emitting region. The estimate of the expected emission intensity within voxel $i$ is $x_i$.

where $q = Ce_N \in \Re^n$ and $e_N \in \Re^N$ is a vector of 1's, so that $q$ is the sum of the columns of $C$ (which need not necessarily be 1). Defining

$$\hat{y} = C^T x$$

to be a *forward transformation,* we can write the gradient and Hessian of the objective function, respectively, as

$$\nabla f_{ML}(x) = -q + C\hat{Y}^{-1}y, \tag{2}$$
$$\nabla^2 f_{ML}(x) = -C\,Y\hat{Y}^{-2}C^T, \tag{3}$$

where $Y = \text{diag}(y_j,\ j = 1,\ldots,N)$ and $\hat{Y} = \text{diag}(\hat{y}_j,\ j = 1,\ldots,N)$. The Hessian is known to be negative semidefinite for this problem (since $y_j/\hat{y}_j \geq 0 \quad \forall j$) and the objective function (1) is concave. Thus, any local maximum will also be a global maximum.

Equation (2) sheds some insight into the computational costs associated with maximizing the objective function. Given a current solution estimate $x^k$, computing the gradient requires first computing a forward transformation $\hat{y}^k = C^T x^k$ and then computing a *backward transformation* $C\hat{Y}_k^{-1}y$ from the forward transformation. The costs of performing the forward transformation and backward transformation are similar and together dominate the computation associated with iterative reconstruction methods, especially in large scale. We shall revisit this computational structure, which is common to all iterative reconstruction methods.

Since the underlying activity distribution is non-negative, the ML reconstruction problem is a constrained optimization problem with lower-bound constraints:

$$\begin{array}{ll} \text{maximize} & f_{ML}(x) \\ \text{subject to} & x \geq 0. \end{array} \tag{4}$$

The ML objective function has a finite maximum and compact level sets on $x \geq 0$ [32].

## 2.1 Maximum *a posteriori* reconstruction

Without regularity conditions on $x$, estimating the spatial emission distribution is a statistically ill-posed problem [6, 29]. The fully converged ML reconstruction, being dominated by noise and edge artifact, is not generally of biomedical interest [49]. Regularization can be included in the objective function by introducing a Bayesian formulation [19, 33]. Given prior probabilities $P\{x\}$ and $P\{y\}$ for the image and measurements, respectively, we define the posterior probability

$$P\{x|y\} = \frac{P\{y|x\}\,P\{x\}}{P\{y\}}.$$

The estimate of $x$ is then obtained by maximizing the posterior probability $P\{x|y\}$.

A common choice for the image prior is the Gibbs distribution $P\{x\} = e^{-\gamma R(x)}$, although other priors (e.g., Gaussian, Gamma) have been investigated [33, 35]. The popularity of Gibbs priors stems in part from their ability to capture the local correlation property of images [18]. The energy function $R$ is defined as a sum of neighborhood functions

$$R(x) = \frac{1}{2}\sum_i \sum_{l\in\mathcal{N}_V} I_{i,l}(x_i, x_l).$$

where $\mathcal{N}_i$ denotes the neighborhood of voxel $i$. In order to maintain concavity and twice continuous differentiability in the objective function, the potential function $V_{i,l}$ is chosen to be convex with continuous first and second derivatives. The potential function is generally designed to discourage non-smoothness in a neighborhood. In our studies we have used the potential function $V_{i,l}(x_i, x_l) = V(x_i - x_l)$, where

$$V(z) = \delta^2\left(\left|\frac{z}{\delta}\right| - \log\left(1 + \left|\frac{z}{\delta}\right|\right)\right) \tag{5}$$

and $\delta$ is a shaping constant that we typically set to 1.

For maximum *a posteriori* (MAP) reconstructions, the objective function (up to a constant) is the log-posterior likelihood

$$f_{MAP}(x) = \log P\{x|y\} = f_{ML}(x) - \gamma R(x). \tag{6}$$

The MAP reconstruction problem can also be posed as a constrained optimization problem

$$\begin{array}{ll} \text{maximize} & f_{MAP}(x) \\ \text{subject to} & x \geq 0. \end{array} \tag{7}$$

We note for future reference the following:

$$\begin{aligned} \nabla f_{MAP}(x) &= \nabla f_{ML}(x) - \gamma\nabla R(x) = -q + C\,\hat{Y}^{-1}y - \gamma\nabla R(x), \\ \nabla^2 f_{MAP}(x) &= \nabla^2 f_{ML}(x) - \gamma\nabla^2 R(x) = -C\,Y\hat{Y}^{-2}C^T - \gamma\nabla^2 R(x). \end{aligned}$$

When a Gibbs prior such as (5) is chosen, the only direction of non-positive curvature of $R$ is uniform field, which is a direction of positive curvature for $f_{ML}$, and therefore $f_{MAP}$ is strictly concave and $\nabla^2 f_{MAP}(x)$ is negative definite [34]. The MAP objective function has a finite maximum and bounded level sets on $x \geq 0$ [33].

## 3 The EM algorithm

The expectation maximization (EM) method, as presented by Dempster, Laird, and Rubin [7] for ML estimation, is an iterative algorithm for computing ML estimates when the measurements are viewed as incomplete data. Shepp and Vardi [47] and Lange and Carson [32] applied the EM method to emission and transmission tomography problems, respectively. The EM algorithm has been proven to converge to an optimal solution of (4) [32, 50].

The EM algorithm requires that a positive initial solution $x^0 > 0$ be specified. Application of the EM method to the unregularized ML reconstruction problem for emission tomography yields the update equation

$$x^{k+1} = \mathcal{M}\left(x^k\right) = X_k Q^{-1} C \hat{Y}_k^{-1} y, \tag{8}$$

where $x^k$ is the current image estimate, $\hat{y}^k = C^T x^k$, and $X_k$, $Q$, and $\hat{Y}_k$ are diagnonal matrices corresponding to the vectors $x^k$, $q$, and $\hat{y}^k$ respectively. The EM algorithm maintains non-negativity at every iteration and converges to a fixed point $x^\infty = \mathcal{M}\left(x^\infty\right)$ which is an optimal solution of (4). The asymptotic rate of convergence is governed by the spectral radius of $\nabla\mathcal{M}\left(x^\infty\right)$ which is typically very close to unity. In one example using reasonable assumptions about the scanner geometry, the lower bound of the spectral radius was calculated to be .99938 [16]. Indeed, EM has been observed to converge very slowly, especially close to the optimal solution. The slow convergence of the EM algorithm has limited its clinical applicability. The cost of one EM iteration is equivalent to the cost of one gradient calculation.

In MAP-EM, the presence of the regularizing term in (6) precludes a closed-form update equation such as (8) for ML-EM. We mention two algorithms that are commonly used for MAP-EM reconstructions: the "one step late" (OSL) algorithm and DePierro's algorithm. Green's OSL algorithm approximates $R\left(x\right)$ with the constant $R\left(x^k\right)$, thereby permitting a closed-form approximated update [15, 16]. OSL converges to the MAP solution provided that $\gamma \leq \bar{\gamma}$, where $\bar{\gamma}$ is an upper threshold for the prior strength. DePierro's algorithm is a "true" MAP-EM implementation that substitutes the convex function $R\left(x\right)$ with a separable, convex, and twice continuously differentiable function $R\left(x, x^k\right) \geq R\left(x\right)$, so that separable maximizations can be performed on the variables [8, 9]. Regularization improves the convergence rate of EM, with larger prior strengths resulting in lower spectral radii. However, for reasonable prior strengths (mild to moderate smoothing), the convergence rates of OSL and DePierro's algorithm are still quite close to unity.

## 4 A primal-dual approach

The known slow convergence of the EM algorithm motivates our investigation into interior-point approaches for the ML and MAP reconstruction problems. As is clear from (1), the objective function can be undefined outside the feasible region $x \geq 0$. Thus the ML and MAP reconstruction problems would appear to be "natural" candidates for interior-point algorithms. Primal-dual methods have enjoyed considerable success in linear programming [17, 28, 36], and have recently been proposed for nonlinear programming [4, 12, 37]. Although they are closely related to the logarithmic barrier method, primal-dual methods may pose some advantages. In the logarithmic barrier method, the Lagrange multiplier estimates may be inaccurate when the primal variables are not close to the barrier trajectory [10]. Primal-dual methods offer the potential of improved "centering" over barrier methods. Given the size of the current problem, the developments presented here must be suitable for large-scale parallel computation.

For convenience of notation, let us repose the reconstruction problem as a constrained *minimization* problem:

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \geq 0. \end{array} \tag{9}$$

Because we are more interested in the fully converged MAP solution than that of ML, we shall henceforth assume that $f(x) = -f_{MAP}(x)$, where $\gamma > 0$. The Karush-Kuhn-Tucker (KKT) first-order necessary conditions for optimality of (9) at a point $x^*$ are existence of Lagrange multipliers $\lambda^*$ so that

$$\nabla \ell(x^*, \lambda^*) = \nabla f(x^*) - \lambda \quad {}^* = \quad 0, \tag{10}$$

$$\lambda_i^* x_i^* \quad = \quad 0, \quad i = 1, \ldots n, \tag{11}$$

$$x^*,\ \lambda^* \quad \geq \quad 0, \tag{12}$$

where $\ell(x, \lambda) = f(x) - \lambda^T x$ is the Lagrangian function. Due to the strict convexity of $f$, the second-order sufficiency conditions are satisfied, and $x^*$ is the unique minimizer of $f$.

In a manner similar to classical barrier methods, primal-dual methods attempt to follow the "barrier trajectory," a smooth trajectory characterized by a *barrier parameter* $\mu > 0$ [11]. The points $(x(\mu), \lambda(\mu))$ along the trajectory satisfy a perturbed version of the KKT conditions:

$$\nabla f(x(\mu)) - \lambda(\mu) \quad = \quad 0, \tag{13}$$

$$\lambda_i(\mu)\, x_i(\mu) \quad = \quad \mu, \quad i = 1, \ldots, n, \tag{14}$$

$$x(\mu),\ \lambda(\mu) \quad > \quad 0. \tag{15}$$

Defining $X = \text{diag}\{x_i, i = 1, \ldots, n\}$ and $\Lambda = \text{diag}\{\lambda_i, i = 1, \ldots, n\}$, our method maintains (15) while attempting to solve (13), (14), that is

$$\begin{bmatrix} \nabla f(x) - \lambda \\ \Lambda X e_n - \mu e_n \end{bmatrix} = 0. \tag{16}$$

Given the point $\left(x^k, \lambda^k\right)$ and the barrier parameter $\mu = \mu_k$, the search direction $p = \left[p_x^T,\ p_\lambda^T\right]^T$ prescribed by Newton's method satisfies the "unsymmetric" primal-dual equations [37]:

$$\begin{bmatrix} \nabla^2 f\left(x^k\right) & -I \\ \Lambda_k & X_k \end{bmatrix} \begin{bmatrix} p_x \\ p_\lambda \end{bmatrix} = - \begin{bmatrix} \nabla f\left(x^k\right) - \lambda^k \\ \Lambda_k X_k e_n - \mu_k e_n \end{bmatrix}. \tag{17}$$

Elimination of the (1,2) block of the matrix in (17) yields the *reduced system*

$$M_k p_x \quad = \quad -\nabla f(x^k) + \mu_k X_k^{-1} e_n, \tag{18}$$

$$p_\lambda \quad = \quad -\lambda^k - X_k^{-1} \Lambda_k p_x + \mu_k X_k^{-1} e_n. \tag{19}$$

where the "condensed" primal-dual matrix is given by

$$M_k = \nabla^2 f(x^k) + X_k^{-1} \Lambda_k. \tag{20}$$

There is no guarantee from (18), (19) that a single steplength taken along the solution $p_x,\ p_\lambda$ maintaining both $x > 0$ and $\lambda > 0$ will be bounded away from zero. We have implemented an algorithm in which the primal and dual variables are permitted to take separate steplengths:

$$\begin{bmatrix} x^{k+1} \\ \lambda^{k+1} \end{bmatrix} = \begin{bmatrix} x^k + \alpha_x p_x \\ \lambda^k + \alpha_\lambda p_\lambda \end{bmatrix}.$$

The primal steplength $\alpha_x$ is chosen to ensure sufficient decrease in the merit function

$$F(x,\mu) = f(x) - \mu \sum_i \log x_i.$$

Observe that $F(x,\mu)$ is simply the logarithmic barrier function and that

$$\nabla_x F(x,\mu) = \nabla f(x) - \mu X^{-1} e_n$$

is identical to the right-hand side of (18) for $\mu = \mu_k$ and $x = x^k$. The unconstrained minimizer $x(\mu)$ of $F(x,\mu)$ satisfies the perturbed KKT conditions (13)-(15) with corresponding multiplier $\lambda_i(\mu) = \mu / x_i(\mu)$, $i = 1, \ldots, n$. Furthermore, the solution of the condensed primal-dual Newton equation (18) is guaranteed to be a descent direction of the merit function for $\mu > 0$, since

$$(\nabla_x F(x,\mu))^T p_x = p_x^T M p_x,$$

and $M$ is positive definite. We shall discuss in further detail the computation of the primal search direction and step length.

The formula for the dual step length follows a suggestion by Conn, Gould, and Toint (CGT) [4]. If $\lambda^{(k)} + p_\lambda$ lies component-wise in the interval

$$\begin{aligned} &\lambda^k + p_\lambda \in \\ &\left[\zeta^{-1} \min\left(e_n, \lambda^k, \mu_k X_{k+1}^{-1} e_n\right), \max\left(\zeta e_n, \lambda^k, \zeta \mu_k^{-1} e_n, \zeta \mu_k X_{k+1}^{-1} e_n\right)\right] \end{aligned} \tag{21}$$

(where $\zeta$ is a constant parameter that we have set to 100) then $\lambda^{k+1} = \lambda^k + p_\lambda$; otherwise find $0 < \alpha_\lambda < 1$ such that $\lambda^{k+1} = \lambda^k + \alpha_\lambda p_\lambda$ minimizes

$$\left\| \Lambda_{k+1} X_{k+1} e_n - \mu_k e_n \right\|_\infty \tag{22}$$

subject to $\lambda^{k+1}$ being in the interval (21). These conditions on the dual step might appear at first glance to be overly restrictive but are actually designed to give maximum flexibility in the choice of $\lambda^{k+1}$. CGT use these bounds on $\lambda$ and nonsingularity of $M$ to prove that, for any fixed parameter value $\bar{\mu}$, the minimization of $F(x,\bar{\mu})$ must be successful, that is, eventually a solution is found that satisfies the perturbed KKT conditions (13)-(15).

In general it is neither necessary nor desirable to reach full subproblem convergence. Rather, we have implemented a "short-step" algorithm in which only one primal-dual step is usually needed before adjusting $\mu$. Setting the barrier parameter $\mu$ is an important consideration in primal-dual algorithms, and has a strong influence on the convergence rate. A reduction in $\mu_k$ is performed whenever the "$\mu$-criticality" conditions [4, 48] are satisfied:

$$\frac{\left(\lambda^{k+1}\right)^T x^{k+1}}{n} \leq \vartheta^C \mu_k, \tag{23}$$

$$\left\| \nabla f\left(x^{k+1}\right) - \lambda^{k+1} \right\|_\infty \leq \vartheta^{DF} \mu_k, \tag{24}$$

where $\vartheta^C$ and $\vartheta^{DF}$ are constant parameters. If the above conditions are satisfied, the barrier parameter is reduced according to

$$\mu_{k+1} = \frac{\left(\lambda^{k+1}\right)^T x^{k+1}}{n\rho}, \tag{25}$$

where $\rho$ is a constant parameter such that

$$\frac{\vartheta^C}{\rho} < 1. \tag{26}$$

A consequence of (26) is that $\mu_k$ cannot increase. Furthermore, since the minimization of $F(x, \mu)$ must be successful, a $\mu$-critical solution (a weaker requirement) must eventually be found. Thus it is impossible for $\mu_k$ to be non-decreasing. Using this argument, CGT prove that the algorithm must converge to a KKT solution [4].

In practice we find that both the primal and dual direction vectors are well scaled, and that $\alpha_x$ and $\alpha_\lambda$ are both typically close to 1. By far the most costly operations are computing the primal direction $p_x$ and updating the gradient $\nabla F(x)$, as we shall explore. In contrast, the costs of the line search for the primal steplength, the computation of the dual search direction (19), and the dual line search (21) are relatively insignificant. From empirical evidence in our computational studies, we have found that a "short-step" algorithm with gradual reduction in $\mu$ achieves the fastest convergence to the KKT conditions. Specifically, we define $\vartheta^C = 1.9$, $\rho = 2$, and $\vartheta^{DF} = 100$. These parameter values enable the $\mu$-critical conditions to be met after only one primal-dual step for most subproblems.

### 4.1 Computing the primal direction

For large problems, factorizing the condensed primal-dual matrix $M$ or even forming the Hessian $\nabla^2 f(x)$ would be prohibitive due to the size of the matrix (376,000×376,000 for even the smaller reconstructions being considered in this paper) and the enormous amount of computation that would be required. Thus we must consider methods for approximating the Newton direction in (18). The approach we have successfully applied to this problem is motivated by the *truncated-Newton* [5] method of unconstrained optimization. The search direction is an approximate or *truncated* solution to the Newton equations [41, 43]

$$Mp_x \approx -\nabla F(x). \tag{27}$$

An early-terminated conjugate gradient (CG) iteration [20] is used to obtain an approximate solution to (27).

An equivalent statement of (27) is we seek to find the direction $p_x$ that approximately minimizes the quadratic $Q(p_x) = \frac{1}{2} p_x^T M p_x + \nabla F(x) p_x$. A reasonable and effective truncation point for (27), based on the monotonicity of $Q(p_x)$, is proposed in [42]; the CG is terminated at subiteration $l$ if

$$\frac{Q\left(p_x^l\right) - Q\left(p_x^{l-1}\right)}{Q\left(p_x^l\right)} \leq \frac{1}{2l}. \tag{28}$$

The CG termination rule (28) has been an important component of the reconstruction software in that it consistently yields a well-scaled primal direction vector as long as $\mu \geq \mu_s$, where $\mu_s$ is a threshold value below which stabilization is required (we shall discuss the $\mu < \mu_s$ case in Section 4.2).

The CG method does not require storage of the Hessian or condensed primal-dual matrix, but rather only application of matrix-vector products. From (3) we can write the first term of the matrix-vector product

$$\nabla^2 f_{ML}(x)\, v = -C\, Y \hat{Y}^{-2} C^T v \tag{29}$$

for an arbitrary vector $v \in \Re^n$. Computationally, (29) consists of a forward transformation ($C^T v$) followed by a diagonal scaling ($\hat{y}$ is already available from the computation of $\nabla f(x)$), followed by a backward transformation (premultiplication by $C$). To be explicit, recalling (20), we have

$$Mv = C\,Y\hat{Y}^{-2}C^T v + \gamma\nabla^2 R(x)\,v + X^{-1}\Lambda v,$$

where $\nabla^2 R(x)\,v$ can be computed exactly without incurring significant computational expense. The forward-and-back-transformation operation in (29) dominates the computational cost of a CG iteration. This operation is computationally similar to computing the gradient, or one EM iteration.

Some authors advocate solving simultaneously for $p_x$ and $p_\lambda$, using the full unsymmetric primal-dual equations (17), or an equivalent symmetrized system [12, 13, 46, 55]. The unsymmetric primal-dual matrix in particular remains nonsingular, and its condition number remains bounded as $\mu \to 0$ [11, 37], when the standard conditions of a constraint qualification, strict complementarity, and the second-order sufficient conditions are satisfied at the solution. Solving the symmetrized system requires significantly more computation ($2n$ variables rather than $n$) without avoiding the potential for ill-conditioning. In our application the Hessian is so large that this is not practical, since we must use inexact solutions. The advantage of using the condensed system (18)–(19) is that although the primal search direction is computed inexactly, the equation for maintaining complementarity (19) is maintained. In practice we find that the resulting primal and dual direction vectors are both well scaled, and that $\alpha_x$ and $\alpha_\lambda$ are typically close to 1.

### 4.1.1 Preconditioning

The use of a preconditioner with the CG is essential for a competitive algorithm. Since every CG subiteration is as costly as a gradient evaluation or EM iteration, it is highly desirable to obtain a quality direction vector in as few CG iterations per subproblem as possible. We have investigated a number of preconditioners, including FFT-based preconditioners that model the approximately Toeplitz-block-Toeplitz nature of $CC^T$ with a circulant-block-circulant approximation [1, 2], high-pass filter approximations to the FFT-based preconditioner [3], the EM preconditioner $XQ^{-1}$ [30], the exact diagonal of $M$, and diagonal Hessian approximations [40]. As a result of the spatially-variant dependence of $M$ on $y$, $\hat{y}$, $x$, and $\lambda$, shift-invariant Toeplitz models of $M$ are highly inappropriate. Of the above preconditioners, by far the best-performing was the exact diagonal of $M$, which can be computed at reasonable cost:

$$M_{i,i} = \sum_j \frac{C_{i,j}^2 y_j}{\hat{y}_j^2} + \gamma\frac{\partial^2}{\partial x_i^2}R(x) + \frac{\lambda_i}{x_i}. \tag{30}$$

Note that the first right-hand side term in (30) is similar in form to a backward transformation, although a bit more expensive due to the squaring operations. We have found that the preconditioned CG method using an exact diagonal preconditioner in the form of (30) almost always requires using fewer than 10 iterations to achieve (28), regardless of the size of the problem. In many cases, only 3 or 4 CG iterations are required. Moreover, the directions produced using an exact diagonal preconditioner are well scaled (usually resulting in primal step sizes of near 1), and lead to rapid descent.

### 4.1.2 Line search

For ML and MAP reconstructions, knowledge of the structure of the objective function can lead to a substantial reduction in the cost of implementing a line search over a more naive approach.

Specifically, after the search direction $p_x$ has been found, and once a forward transformation $\hat{w} = C^T p_x \in \Re^N$ has been computed, it is possible to compute the objective function and first and second directional derivative values at the trial points $\left(x^k + \alpha p_x\right)$ at nearly negligible cost. To see this, note that $\hat{y}^{k+1} = \hat{y}^k + \alpha\hat{w}$, and therefore [26, 27]

$$f\left(x^k + \alpha_x p_x\right) = q^T x^k + \alpha_x q^T p_x - \sum_j y_j \log\left(\hat{y}_j^k + \alpha_x \hat{w}_j\right) + \gamma R\left(x^k + \alpha_x p_x\right).$$

Similar expressions exist for the directional first and second derivatives [23].

After the initial forward transformation to compute $\hat{w}$, no further forward- or back-transformation operations are required during the line search at any of the trial points. The forward transformation $\hat{w}$ can be re-used, so that only one backward transformation is subsequently required to update the gradient. The above observations and the well behaved convex nature of the objective function have permitted us to implement a highly accurate but low-cost Newton line search. Due to the low cost of each step we have chosen a relatively strict tolerance of 0.05 on the Wolfe condition for termination of the line search. We find this line search technique to be highly effective and, in no small part, responsible for the positive results we report.

### 4.2 Stabilization

A well known property of the Hessian of the primal barrier function is its increasingly ill-conditioned nature as $\mu \to 0$ [39]. Analogous results hold for the condensed primal-dual matrix: as the solution is approached the matrix becomes increasingly ill-conditioned. (For a detailed analysis see the paper by Wright [54]).

In [44], Nash and Sofer developed an approximation to the Newton direction for the logarithmic barrier, that avoids the structural ill-conditioning of the barrier Hessian and is suitable for large-scale problems. The direction is the sum of two vectors, one in the null space of the Jacobian of the active constraints, and the other orthogonal to it. The associated decoupling is based on a prediction of the binding set at the solution.

We have recently adapted this approximation to the condensed Newton equations arising in primal-dual methods. Although our derivation is valid for general nonlinear constraints, we present it here for the special case of bound constraints in the context of (18).

We will assume in the following that strict complementarity holds at the solution, that is, $\lambda_i^* > 0$ if $x_i^* > 0$. Define $\mathcal{J} = \{i : x_i^* = 0\}$ to be the index set of binding constraints at the solution, and $\hat{n}$ to be the number of binding constraints at the solution. We will assume that $0 < \hat{n} < n$, as is always the case in reconstructions of practical interest. Define $\mathcal{I} = \{i : x_i^* > 0\}$ the set of nonbinding constraints. Let $x^{\mathcal{I}}$ be the subvector of variables that are positive at the optimal solution, and $x^{\mathcal{J}}$ the subvector of variables that are zero at the optimal solution. Assume also that the variables are ordered so that the positive variables are first, i.e.,

$$x = \begin{bmatrix} x^{\mathcal{I}} \\ x^{\mathcal{J}} \end{bmatrix},$$

The Hessian of the objective function $H = \nabla^2 f(x)$ will then be similarly partitioned, as will the condensed primal-dual matrix

$$M = \begin{bmatrix} M_{\mathcal{I},\mathcal{I}} & M_{\mathcal{I},\mathcal{J}} \\ M_{\mathcal{I},\mathcal{J}}^T & M_{\mathcal{J},\mathcal{J}} \end{bmatrix} = \begin{bmatrix} H_{\mathcal{I},\mathcal{I}} + X_{\mathcal{I}}^{-1}\Lambda_{\mathcal{I}} & H_{\mathcal{I},\mathcal{J}} \\ H_{\mathcal{I},\mathcal{J}}^T & H_{\mathcal{J},\mathcal{J}} + X_{\mathcal{J}}^{-1}\Lambda_{\mathcal{J}} \end{bmatrix},$$

where $X_{\mathcal{I}}$, $X_{\mathcal{J}}$, $\Lambda_{\mathcal{I}}$ and $\Lambda_{\mathcal{J}}$ are the diagonal matrices of the associated components of $x$ and $\lambda$.

We will assume that the sequence of iterates $(x, \lambda)$ generated by the primal-dual satisfies the following properties, when $\mu$ is sufficently small:

$$\begin{aligned} x_i &= \Theta(1), \quad \lambda_i = \Theta(\mu), \quad i \in \mathcal{I}, \\ x_i &= \Theta(\mu), \quad \lambda_i = \Theta(1), \quad i \in \mathcal{J}. \end{aligned}$$

Here we define $\tau = \Theta(\mu)$ if there exist constants $0 < \kappa_l < \kappa_u$ so that $\kappa_l \mu \leq \|\tau\| \leq \kappa_u \mu$ for all sufficiently small $\mu > 0$. We say that a vector or matrix is $\Theta(\mu)$ if its norm is $\Theta(\mu)$. We also define $\tau = O(\mu)$ if there exists some positive constant $\kappa_u$ so that $\|\tau\| \leq \kappa_u \mu$ for all sufficiently small $\mu > 0$.

We will also assume that near the solution the Hessian is reasonably well conditioned, so that $H = O(1)$. Now the diagonal terms of $M_{\mathcal{J},\mathcal{J}}$ are $O(1/\mu)$, and become unbounded as $\mu \to 0$. In contrast, the diagonal terms of $M_{\mathcal{I},\mathcal{I}}$ differ from those of the reduced Hessian $H_{\mathcal{I},\mathcal{I}}$ by $O(\mu)$, and the condition of $M_{\mathcal{I},\mathcal{I}}$ thus reflects that of the constrained problem. The condensed primal-dual matrix $M$ can then be shown to have $\hat{n}$ "large" eigenvalues of magnitude $\Theta(1/\mu)$, and $n - \hat{n}$ "small" eigenvalues that differ from those of $H_{\mathcal{I},\mathcal{I}}$ by $O(\mu)$, and have magnitude $\Theta(1)$. The condensed primal-dual matrix thus suffers from the same structured ill-conditioning as the barrier Hessian.

For small values of $\mu$ we propose approximating the primal Newton direction $p_x$, by a direction $\tilde{p}_x$, whose null- and range-space components are computed as follows:

$$(M_{\mathcal{I},\mathcal{I}})\, \tilde{p}_x^{\mathcal{I}} = -\left(\nabla F^{\mathcal{I}} - M_{\mathcal{I},\mathcal{J}} X_{\mathcal{J}} \Lambda_{\mathcal{J}}^{-1} \nabla F^{\mathcal{J}}\right), \tag{31}$$

$$\tilde{p}_x^{\mathcal{J}} = -X_{\mathcal{J}} \Lambda_{\mathcal{J}}^{-1} \nabla F^{\mathcal{J}}. \tag{32}$$

The system for computing the component $\tilde{p}_x^{\mathcal{I}}$ involves the well conditioned matrix $M_{\mathcal{I},\mathcal{I}}$, and can be solved exactly or inexactly via the conjugate gradient method. The computation of $\tilde{p}_x^{\mathcal{J}}$ is straightforward. Thus, the ill-conditioning of the condensed primal-dual is avoided. We will show now that under the assumptions above, $\tilde{p}_x - p_x = O(\mu^2)$, so that the accuracy of the approximation increases as the solution is approached and the potential harm from ill-conditioning increases.

Using the well known formula for the inverse of a partioned matrix (see e.g. [45, 55]) it follows that

$$\begin{aligned} \tilde{p}_x^{\mathcal{I}} - p_x^{\mathcal{I}} &= (M_{\mathcal{I},\mathcal{I}})^{-1} \left( M_{\mathcal{I},\mathcal{J}} G^{-1} M_{\mathcal{I},\mathcal{J}}^T M_{\mathcal{I},\mathcal{I}}^{-1} \nabla F^{\mathcal{I}} + M_{\mathcal{I},\mathcal{J}} (X_{\mathcal{J}} \Lambda_{\mathcal{J}}^{-1} - G^{-1}) \nabla F^{\mathcal{J}} \right), \\ \tilde{p}_x^{\mathcal{J}} - p_x^{\mathcal{J}} &= -G^{-1} M_{\mathcal{I},\mathcal{J}}^T M_{\mathcal{I},\mathcal{I}}^{-1} \nabla F^{\mathcal{I}} - (X_{\mathcal{J}} \Lambda_{\mathcal{J}}^{-1} - G^{-1}) \nabla F^{\mathcal{J}}, \end{aligned}$$

where

$$G = M_{\mathcal{J},\mathcal{J}} - M_{\mathcal{I},\mathcal{J}}^T M_{\mathcal{I},\mathcal{I}}^{-1} M_{\mathcal{I},\mathcal{J}}.$$

Now since $M_{\mathcal{J},\mathcal{J}} = X_{\mathcal{J}}^{-1} \Lambda_{\mathcal{J}} + H_{\mathcal{J},\mathcal{J}}$,

$$G = X_{\mathcal{J}}^{-1} \Lambda_{\mathcal{J}} \left( I + \Lambda_{\mathcal{J}}^{-1} X_{\mathcal{J}} O(1) \right) = X_{\mathcal{J}}^{-1} \Lambda_{\mathcal{J}} \left( I + O(\mu) \right),$$

so that $G^{-1} = O(\mu)$ and

$$G^{-1} - X_{\mathcal{J}} \Lambda_{\mathcal{J}}^{-1} = O(\mu^2).$$

Note further, that

$$\nabla F^{\mathcal{I}} = (\nabla f(x))^{\mathcal{I}} - \mu X_{\mathcal{I}}^{-1} e = O(\mu) + \mu O(1) = O(\mu),$$

whereas

$$\nabla F^{\mathcal{J}} = (\nabla f(x))^{\mathcal{J}} - \mu X_{\mathcal{J}}^{-1} e = O(1) + \mu O(1/\mu) = O(1).$$

It follows that

$$\tilde{p}_x^{\mathcal{I}} - p_x^{\mathcal{I}} = O(\mu) O(\mu) + O(\mu^2) = O(\mu^2),$$

Table 1: An example of the effect of stabilization. The number of CG iterations, *ncg*, is counted from the beginning of the $\mu = 1.40 \times 10^{-4}$ subproblem. The termination condition in this example is $\lambda^T x/n \leq 7.5 \times 10^{-5}$.

| non-stabilized | | | | stabilized | | | |
|---|---|---|---|---|---|---|---|
| $\mu$ | $\alpha_x$ | $\alpha_\lambda$ | *ncg* | $\mu$ | $\alpha_x$ | $\alpha_\lambda$ | *ncg* |
| 1.49E-4 | 0.950 | 0.003 | 5 | 1.49E-4 | 0.942 | 1.000 | 12 |
| 1.26E-4 | 0.156 | 1.000 | 17 | 1.16E-4 | 0.923 | 1.000 | 17 |
| 1.01E-4 | 1.000 | 0.002 | 22 | 7.75E-5 | 0.159 | 0.8245 | 30 |
| 8.50E-5 | 0.143 | 1.000 | 34 | 6.47E-5 | 0.962 | 1.000 | 35 |
| 7.08E-5 | 0.392 | 1.000 | 46 | 3.25E-5 | | | |
| 5.83E-5 | 1.000 | 0.166 | 51 | | | | |
| 4.77E-5 | 0.016 | 1.000 | 62 | | | | |
| 3.75E-5 | | | | | | | |

and

$$\tilde{p}_x^{\mathcal{J}} - p_x^{\mathcal{J}} = O(\mu)O(\mu) + O(\mu^2) = O(\mu^2),$$

so that $\tilde{p}_x - p_x = O(\mu^2)$.

In [44], Nash and Sofer prove (for the case of the Newton direction arising from the logarithmic barrier objective function) that, for sufficiently small $\mu$, the vector computed using an approximation similar to (31) and (32) yields a descent direction with respect to the logarithmic barrier objective function. The proof is readily extended to the present primal-dual case; thus $p_x$ is a descent direction for the merit function $F(x, \mu)$. We have found that, for the present problem, the above approximation to the Newton direction is useful for values of $\mu$ of order $10^{-4}$ or less.

Recently [54] showed that the errors generated by backward-stable numerical methods (various Cholesky factorizations and Gaussian elimination with partial pivoting) for solving (18) are not magnified by the structured ill-conditioning. These methods are inappropriate for our large problems which involve potentially millions of variables. Instead we find an approximate solution using a CG iteration. When working in inexact arithmetic with large numbers of variables, the convergence rate of the CG method depends on the condition of $M$ [14]. Thus the structural ill-conditioning in $M$ can lead the CG iteration to spend an unnecessary amount of work in computing $p_x$. Further, as we have observed, the criterion for terminating the CG may be overly optimistic in an ill-conditioned system, so that the resulting direction is poorly scaled as $\mu \to 0$.

The potential effect of ill-conditioning is illustrated through an example in Table 1. This example was encountered during development and motivated the incorporation of stabilization into the algorithm. Starting at the subproblem $\mu = 1.49 \times 10^{-4}$, the primal steplength, dual steplength, and *ncg* (the number of CG iterations), are listed for both the non-stabilized and stabilized cases. This test was terminated at $\lambda^T x/n \leq 7.5 \times 10^{-5}$. Note that in the non-stabilized case, the number of CG iterations from the first subproblem in the test to termination is significantly lower in the stabilized test than the non-stabilized test. Note also that in many of the non-stabilized subproblems, either the primal or dual steplength is small, indicating a poorly scaled direction or loss of accuracy.

There has been much recent interest in stabilization methods that do not require a prediction of the active set [12, 13, 53]. These approaches are based on factorization methods which are unsuitable for a problem as large as the present one. The argument against stabilization methods that require a prediction set is that the active set is unknown in interior-point methods. We argue that, close to the solution in the emission tomography reconstruction problem, an accurate prediction of the active set can be made. In our problem, the constraints have a simple interpretation. The positive variables correspond to those voxels containing at least some radioactive tracer, while the zero-

valued variables correspond to those voxels that lack any tracer activity. Close to the solution, when $\mu$ becomes sufficiently small that stabilization is appropriate, the set of binding constraints is obvious and can be conservatively identified with a $\mu$-dependent threshold.

### 4.3 Extrapolation

Fiacco and McCormick showed that the solutions $x(\mu)$ at the perturbed KKT solutions form a unique differentiable trajectory in $\mu$ [11]. The perturbed KKT conditions (13)–(15) define a "central path" as $\mu \to 0$. Thus, a successful algorithm may be able to move both "along" and "toward" the path. As discussed in [11], from the subproblem solutions $\{x(\mu_l),\ l = 1 \ldots k\}$ the trajectory can be approximated as a polynomial

$$x(\mu) \simeq \sum_{l=k-r}^{k} c_l \mu^l, \tag{33}$$

where $r$ is the degree of the approximating polynomial and $c_{k-r}, \ldots, c_k$ are $r+1$ vectors of coefficients. Using the approximation in (33), we find a direction $\Delta x$ such that

$$\Delta x = \sum_{l=k-r}^{k} c_l \mu^l - x^k,$$

and set

$$\hat{x}(\mu_{k+1}) = x^k + \bar{\alpha}\ \Delta x \tag{34}$$

to be a prediction to the next subproblem's primal solution. Here $x^k$ is the computed (approximate) subproblem solution for $\mu = \mu_k$. Primal feasibility is maintained by the steplength $\bar{\alpha} = 0.98\alpha_{\max}$, where $\alpha_{\max}$ is the maximum steplength that does not violate non-negativity in $x$. Then, in the manner of (19), we compute a dual direction vector according to

$$\Delta\lambda\ = -\lambda^k - X_k^{-1}\Lambda_k(\bar{\alpha}\triangle x) + \mu X_k^{-1} e_n. \tag{35}$$

The dual vector is then moved according to

$$\hat{\lambda}(\mu_{k+1}) = \lambda_k + \tilde{\alpha}_\lambda \triangle x, \tag{36}$$

which requires another dual line search to minimize (22). The predicted solution $\left(\hat{x}(\mu_{k+1}), \hat{\lambda}(\mu_{k+1})\right)$ then serves as a starting point for the $(k+1)$st subproblem, a prediction to the solution at $\mu_{k+1}$. The extrapolated primal-dual method can be viewed as a predictor-corrector algorithm, with the extrapolation (34 and 36) serving as the "predictor" step, and the subproblem minimization serving as the centering or "corrector" step [22]. The degree $r$ of the approximating polynomial is 1 when predicting the 3rd subproblem, 2 for the 4th, and 3 for the 5th and beyond.

We have experimented with line searches in conjunction with (34), but often $\bar{\alpha} \ll 1$, and hence the line search just yields $\bar{\alpha}$. For this reason, we have found that (36) yields a more effective dual direction than does the equivalent of (19) in the context of extrapolation. Although the extrapolated search direction $\Delta x$ can often be poorly scaled (i.e., $\bar{\alpha} \ll 1$), we have observed that the directions produced are always descent directions to the merit function and lead to a significant decrease in the objective function $f$. A number of reconstructions were performed in which $\Delta\lambda$ was computed by extrapolating the dual solution vector (rather than computing it via (36)); the discouraging nature of the results led us to abandon direct extrapolation of the dual vector in favor of (36) which is highly effective in comparison.

Following extrapolation, an additional gradient evaluation $\nabla F\left(\hat{x}\left(\mu_{k+1}\right), \mu_{k+1}\right)$ is required. Since the primal-dual algorithm requires between 12 and 25 subproblems to perform a 3-D MAP reconstruction, extrapolation adds that many gradient evaluation operations to the computational cost. So extrapolation is only economical if it reduces the computational burden by at least as much as it adds. Our experience has been that for some data sets, the cost of extrapolation is well worthwhile but for other data sets the benefits were only marginal. Extrapolation thus appears to serve as somewhat of a safeguard against difficult problems. In an extrapolated primal-dual reconstruction, the convergence measure $\max\left(\lambda_i x_i\right)$ does not decrease as monotonically as in a primal dual reconstruction without extrapolation. Certain extrapolated steps seem to cause the algorithm to "get ahead of itself," but this effect is transient. On the studies we've performed, the algorithm does ultimately converge to an accurate solution with extrapolation.

### 4.4 Initialization

The choice of the initial barrier parameter may have a substantial effect on the algorithm. If the parameter is too small, the first subproblem may have extreme difficulty due to ill conditioning; if the parameter is too large, then many (unnecessary) subproblems will be required to solve the problem. Proper initialization of the barrier parameter $\mu$ involves finding the most suitable point on the barrier trajectory based on the initial solution $x^o$ and the measurement data $y$. Recalling the perturbed necessary conditions in (13), if the initial solution $\hat{x}^0$were to be *on* the central path, it would satisfy

$$\nabla F\left(\hat{x}^0, \mu_0\right) = q - C\hat{Y}^{-1}y + \gamma\nabla R\left(\hat{x}^0\right) - \mu_0\hat{X}_0^{-1}e_n = 0.$$

Pre-multiplying by $\left(\hat{x}^0\right)^T$ we arrive at

$$q^T\hat{x}^0 - y^Te_N + \gamma\nabla^T R\left(\hat{x}^0\right)\hat{x}^0 = n\mu_0.$$

This suggests the following rule for initialization, which we find quite effective:

$$\mu_0 = \frac{\left|q^T\hat{x}^0 - y^Te_N + \gamma\nabla^T R\left(\hat{x}^0\right)\hat{x}^0\right|}{n}. \tag{37}$$

Another, similar, initialization rule is motivated by the goal of finding an initial value $\mu_0$ so that

$$\nabla f\left(\hat{x}^0\right) - \mu_0\hat{X}_0^{-1}e_n \approx 0. \tag{38}$$

While (38) cannot be solved exactly, we can try to find a $\mu_0$ that results in a point $\hat{x}^0$ that is close to the barrier trajectory according to, say, the 2-norm. This motivation leads to an alternative initialization rule [45]

$$\mu_0 = \frac{\left\|\nabla f\left(\hat{x}^0\right)\right\|_2}{\left\|\hat{X}_0^{-1}e_n\right\|_2}. \tag{39}$$

During the course of development, both initialization rules were tried on certain data sets. Although both initialization rules performed well, reconstructions initialized with (37) usually reached the optimal solution in slightly less overall work than those initialized with (39).

The initial estimate for $\hat{x}^0$ and $\hat{\lambda}^0$ we used most frequently was in each case a positive uniform field. A discussion on the rationale of using a uniform field for $\hat{x}^0$ and on criteria for choosing the constant value of the primal initial solution may be found in [23]. Alternative choices for the initial dual vector may be preferable, and an investigation into this question may be worthwhile.

Table 2: Properties affecting computation, memory, and storage costs for two different-sized reconstruction problems. Gradient evaluation costs are based on a 2.5M-count study on 10 120-MHz IBM RISC/6000 SP processors.

| size class | $n$ | $N$ | elements in $C$ | density in $C$ | storage cost of $C$ | cost of gradient |
|---|---|---|---|---|---|---|
| thick-slice | 376,882 | $5.36 \times 10^6$ | $2.02 \times 10^{12}$ | 0.93% | 390 MB | 3.42 min. |
| thin-slice | $1.40 \times 10^6$ | $6.30 \times 10^7$ | $8.82 \times 10^{13}$ | 0.35% | 1.42 GB | 7.23 min. |

### 4.5 Termination

Given that subproblem termination is based on the $\mu$-criticality conditions (23) and (24), the closeness of each subproblem solution can be measured by $\mu$. It can be shown that if subproblems are solved exactly, $|f(x(\mu)) - f(x^*)| \leq n\mu$ [11]. The $\mu$-criticality conditions, however, are designed for a "short-step" algorithm in which one truncated-Newton step should satisfy each subproblem for sufficiently small $\mu$. To ensure the accuracy of the final solution, final termination is based on the following two requirements:

$$\frac{\lambda^T x}{n} \leq \varepsilon_1, \tag{40}$$

$$\frac{\|\nabla_x \ell(x, \lambda)\|_\infty}{1 + |f(x)|} \leq \varepsilon_2. \tag{41}$$

We have found that reasonably accurate solutions are ensured when $\varepsilon_1 = 1.5 \times 10^{-4}$ and $\varepsilon_2 = 5 \times 10^{-9}$.

The traditional view in tomographic reconstruction is that a highly accurate solution is unnecessary. This view stems in part from the ill-posedness of the problem and the computational cost of taking a reconstruction to full convergence. From empirical evidence in our studies, the ability to perform certain imaging tasks such as "cold spot detectability" improves with accuracy of the solution. Although the termination criteria we propose above may not appear particularly strict, they are from a tomographic reconstruction perspective.

## 5 Computational studies

### 5.1 Size of the problem

Our studies involved two different-sized problems. Raw coincidence data from the scanner can be binned into either "thick-slice" or "thin-slice" measurement spaces, or both. "Thick-slice" reconstructions, in which $n = 376,000$ and $N = 5.35 \times 10^6$, require 3.4 minutes for a gradient evaluation using 10 IBM RISC/6000 SP processors (120 MHz) on a 2.5M-count study. For a "thin-slice" reconstruction with $n = 1.4 \times 10^6$ and $N = 6.3 \times 10^7$ on the same data and processors, a gradient evaluation requires 6.75 minutes. These properties are summarized in Table 2. The cost of storing the full $n \times N$ system matrix is prohibitive, even for thick-slice reconstructions. Extensive exploitation of the sparsity and symmetries inherent in the system matrix makes its storage and retrieval possible [23, 24].

The dominant computational operations of the reconstruction problems are the forward- and back-transformation operations that underlie EM iterations, gradient evaluations, Hessian-vector products, and diagonal Hessian calculations. These operations have been implemented in parallel via a data decomposition strategy that partitions the "measurement-space" vectors $y$ and $\hat{y}$ across

Table 3: Summary of thick-slice primal-dual results and comparison with MAP-EM. Extrapolation was not used, and in all cases $\rho = 2$.

| study | $f^*$ | *npr* | *nit* | *ncg* | *ngr* | MAP-EM |
|---|---|---|---|---|---|---|
| **A** | 2,465,770 | 19 | 19 | 110 | 148 | 1000 |
| **B** | 2,397,197 | 23 | 23 | 164 | 210 | > 1000 |
| **C** | 2,269,180 | 22 | 22 | 126 | 170 | 990 |
| **D** | 2,752,484 | 20 | 21 | 169 | 211 | > 1000 |
| **E** | 2,536,110 | 26 | 26 | 131 | 183 | 770 |
| **F** | 3,296,013 | 23 | 23 | 141 | 187 | >1000 |
| **G** | 3,660,344 | 24 | 24 | 127 | 175 | > 1000 |
| average *ngr* | | | | | 183 | |

the processors. The "image-space" vectors such as $x$ and $\lambda$ are replicated over all processors. Our data decomposition is justifiable under the observation that $N >> n$. On a data set with 2.5M counts, at most 47% of the elements of $y$ will be nonzero in the thick-slice case; at most 4% in the thin-slice case. The dominant computational operations have been implemented in such a way to exploit sparsity in $y$ and further conserve computation [23].

### 5.2 Cost metrics

We have devised metrics to measure the cost of an interior point reconstruction. Define the number of subproblems to be *npr*, the number of truncated-Newton iterations *nit*, the number of conjugate gradient subiterations *ncg*. The cost of one CG iteration (dominated by the Hessian-vector product) is equivalent to the cost of one gradient calculation or EM iteration. One truncated-Newton iteration requires, in addition to the *ncg* operations, one diagonal Hessian evaluation plus one forward transformation and one backward transformation. The exact cost of these operations varies depending on the size of the problem and number of counts, but we shall approximate the cost of one truncated-Newton iteration to be the equivalent of two gradient calculations beyond the cost of the conjugate gradients.

Using this approximation, the total cost of unextrapolated interior-point reconstructions can be measured in units of equivalent number of gradient calculations (or EM iterations):

$$ngr = 2 \cdot nit + ncg.$$

Extrapolation requires an additional gradient calculation following the extrapolation in order to update the gradient vector. With extrapolation we modify the formula to

$$ngr = npr + 2 \cdot nit + ncg.$$

### 5.3 Computational results

We have performed a number of 3-D reconstructions on data acquired from a small animal scanner and data generated by a Monte Carlo simulation of the same small animal scanner. Reconstructions of seven datasets were taken to full convergence, as defined by the termination criteria (40) and (41) with $\epsilon_1 = 1.5 \times 10^{-4}$ and $\epsilon_2 = 5 \times 10^{-9}$. The various datasets used in our computational studies represent a fairly diverse sample of the types of scans that might be encountered in practice. The number of counts in the datasets used in these studies ranged from 850K to 5.1M. The number of binding constraints at the optimal solution ranged from approximately 20% to 80%.

Our main results are summarized in Tables 3 and 4 for the non-extrapolated and extrapolated primal-dual cases, respectively. Studies **A** through **D** are reconstructions of data acquired from a

Table 4: Summary of thick-slice extrapolated primal-dual results and comparison with MAP-EM; in all cases $\rho = 2$.

| study | $f^*$ | *npr* | *nit* | *ncg* | *ngr* | MAP-EM |
|---|---|---|---|---|---|---|
| **A** | 2,465,772 | 17 | 17 | 94 | 145 | 960 |
| **B** | 2,397,232 | 16 | 16 | 91 | 139 | > 1000 |
| **C** | 2,269,190 | 17 | 17 | 94 | 145 | 850 |
| **D** | 2,752,502 | 14 | 16 | 119 | 165 | > 1000 |
| **E** | 2,536,112 | 20 | 20 | 106 | 166 | 750 |
| **F** | 3,296,029 | 18 | 18 | 115 | 169 | >1000 |
| **G** | 3,660,384 | 20 | 20 | 100 | 160 | >1000 |
| average *ngr* | | | | | 156 | |

Table 5: Summary of thick-slice barrier results and comparison with MAP-EM. Extrapolation was used on all data sets, and in all cases $\rho = 10$.

| study | $f^*$ | *npr* | *nit* | *ncg* | *ngr* | MAP-EM |
|---|---|---|---|---|---|---|
| **A** | 2,465,832 | 5 | 28 | 159 | 218 | 880 |
| **B** | 2,397,199 | 5 | 29 | 198 | 259 | > 1000 |
| **C** | 2,269,180 | 5 | 29 | 185 | 246 | 990 |
| **D** | 2,752,499 | 4 | 25 | 207 | 260 | > 1000 |
| **E** | 2,536,111 | 6 | 40 | 214 | 298 | 780 |
| **F** | 3,296,037 | 5 | 36 | 197 | 274 | >1000 |
| **G** | 3,660,351 | 6 | 41 | 214 | 300 | > 1000 |
| average *ngr* | | | | | 265 | |

small animal PET scanner, while studies **E** through **G** are reconstructions of Monte Carlo simulated data. These reconstructions were performed in "thick-slice" mode (376,832 variables) with the regularization parameter set at $\gamma = 3 \times 10^{-4}$. In these tables, the rightmost column indicates the number of DePierro MAP-EM iterations that were required to achieve the value of $f^*$ in the same row. (To avoid excessive computation, the function values were only calculated every 10 MAP-EM iterations, and the final count was rounded down, to favor MAP-EM.) Since the cost of one gradient evaluation is equivalent to the cost of one EM iteration, the numbers in the columns *ngr* and MAP-EM can be compared directly. We find that the primal-dual method consistently reaches convergence much more rapidly than does MAP-EM.

Another interesting observation can be made in the comparison between Tables 3 and 4. Consider the number of EM iterations required to reach $f^*$ for study **C**. In Table 3, the EM algorithm reached $f = 2,269,180$ in 990 iterations. In Table 4 on the same data set, the EM algorithm reached $f = 2,269,190$ in 850 iterations. Thus, the EM algorithm took 140 iterations to reduce the function value by only 10 units near the solution. This is in fact a typical example of the slow limit behavior of the EM algorithm. In all studies, MAP-EM did not achieve the same convergence results obtained by the interior-point methods at termination. The Lagrangian gradient norm and complementary slackness values of the terminated MAP-EM iterates were consistently much higher than those of the terminated primal-dual solution.

We have also performed these reconstructions using a stabilized logarithmic barrier algorithm based on the method presented in [44] and specialized to the present reconstruction problem. Many of the computational features of our logarithmic barrier implementation are identical to our primal-dual implementation, e.g., truncated Newton, line search, computation of the gradient, Hessian-vector product, etc. For a more detailed discussion, see [23]. The logarithmic barrier results are summarized and compared against MAP-EM in Table 5. Termination of the logarithmic barrier

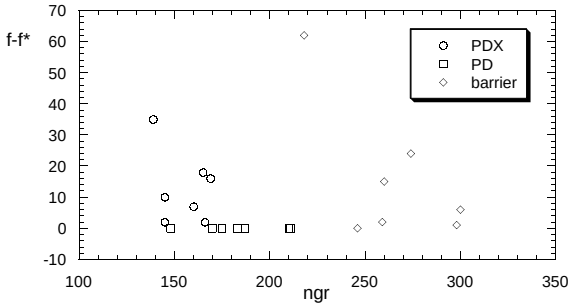


Figure 2: "Distance" from optimal solution at termination, as measured by difference in objective function $f - f^*$ (where $f^*$ is here defined to be the lowest objective function obtained per study), versus work required to reach termination, as measured by $ngr$, the equivalent number of gradient evaluations. The studies included are those listed in Table 3. PD stands for non-extrapolated primal-dual, PDX for extrapolated primal-dual.

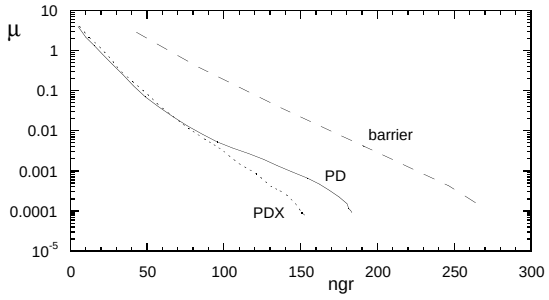


Figure 3: Average value of $\mu$ at subproblem termination versus average $ngr$ (equivalent number of gradient evaluations) for the seven studies listed in Table 3. PD stands for non-extrapolated prmal-dual, PDX for extrapolated primal-dual.

was defined by (41) and

$$\frac{\max\left(\lambda_i x_i\right)}{1 + \left|f\left(x\right)\right|} \leq 5 \times 10^{-10}.$$

These termination criteria for the logarithmic barrier correspond to roughly the same accuracy as (40) and (41) do for the primal-dual method. Being a "long-step" method, the logarithmic barrier gives the user less control over the exact stopping point than does the "short-step" primal-dual. All of the logarithmic barrier reconstructions in Table 5 used extrapolation. In all logarithmic barrier reconstructions, $\mu$ was reduced by a factor of 10 between subproblems. Table 5 indicates that the primal barrier method requires on average 70% more gradient calculations than the primal-dual method on these problems.

The effect of extrapolation is illustrated in Figures 2 and 3. In Figure 2, the equivalent number of gradient evaluations ($ngr$) to reach termination is plotted against objective function "distance" $f - f^*$, the difference between the function value of the terminated solution and the lowest function value obtained for that reconstruction. In all seven test cases (those listed in Tables 3–5), the

Table 6: Summary of thin-slice extrapolated results, including convergence measures and computational costs to optimal solution.

| study | $\gamma$ | $f$ | $\frac{\|\nabla \ell\|}{1+\|f\|}$ | $\frac{\max(\lambda_i x_i)}{1+\|f\|}$ | $npr$ | $nit$ | $ncg$ | $ngr$ |
|---|---|---|---|---|---|---|---|---|
| **B** | 8E-5 | 7,661,605 | 8.54E-11 | 1.77E-11 | 17 | 17 | 68 | 119 |
| **B** | 3E-5 | 7,658,720 | 9.77E-11 | 1.68E-11 | 17 | 17 | 72 | 123 |
| **B** | 1E-5 | 7,657,020 | 3.71E-10 | 2.14E-11 | 17 | 17 | 79 | 130 |
| **C** | 3E-5 | 5,826,032 | 9.08E-10 | 2.67E-11 | 15 | 16 | 56 | 103 |
| **F** | 3E-5 | 7,724,731 | 6.87E-10 | 7.96E-11 | 16 | 16 | 64 | 112 |
| **F** | 1E-5 | 7,721,001 | 1.29E-9 | 2.87E-11 | 14 | 14 | 71 | 113 |
| **H** | 3E-5 | 3,776,745 | 1.10E-9 | 3.89E-11 | 13 | 16 | 63 | 108 |
| average $ngr$ | | | | | | | | 115 |

unextrapolated primal-dual method achieved the lowest objective function value. Thus, $f - f^*$ is zero for all unextrapolated primal-dual (PD) results but greater than zero for the extrapolated primal-dual (PDX) and barrier results. The PDX results are clustered in a region of lower $ngr$ than the PD results. This indicates that extrapolation lowers the computational expense to the solution at a slight deterioration in the final objective. Compared with the barrier method, either extrapolated or unextrapolated primal-dual produces equivalent or better accuracy with less computation required.

In Figure 3, the average number of equivalent gradient evaluations at subproblem termination is plotted against the average value of $\mu$ for each subproblem. Both averages ($ngr$ and $\mu$) were taken from the same seven test cases of Tables 3–5. Compared with either unextrapolated primal-dual (PD) or extrapolated primal-dual (PDX), the logarithmic barrier is clearly on a slower trajectory. The PD and PDX trajectories are quite similar until approximately $\mu = 0.01$, at which point the PD curve "swings out", while the PDX curve continues to descent log-linearly. This result confirms that the prediction (extrapolation) step becomes more accurate near the solution, resulting in more rapid convergence. However, a comparison of the objective functions indicates that the value of PDX $\mu$ is perhaps one step "ahead of itself," compared with the unextrapolated case.

We have also reconstructed a number of very large-scale "thin-slice" reconstructions involving $1.4 \times 10^6$ variables. Table 6 summarizes a number of properties of these extrapolated primal-dual reconstructions at the converged solution. A smaller group of datasets (the more visually "interesting" studies) were selected for the thin-slice work, and certain reconstructions were repeated with different values of the prior strength $\gamma$. Thin-slice reconstructions seem to require a lower prior strength than the corresponding thick-slice reconstructions. The most visually pleasing results were from reconstructions using $\gamma = 1 \times 10^{-5}$, which is 1/30 the prior strength that was generally found to be most satisfactory in thick-slice reconstructions. The total amount of work (as measured in $ngr$) required to reach termination in Table 6 is also quite pleasing. The number of variables in a thin-slice reconstruction is approximately 3.7 times the number in thick-slice. The number of nonzero-valued measurements in thin-slice mode is only marginally greater than in thick-slice mode, however, since the number of counts is the same in both cases. These thin-slice reconstructions may thus be better conditioned than their thick-slice counterparts.

# 6 Conclusion

From the results of the previous Section, it is clear that the primal-dual method converges significantly faster than the EM algorithm for regularized ML reconstructions in emission tomography. The results also indicate that the primal-dual method converges faster than the logarithmic barrier

method. The use of extrapolation in conjunction with the primal-dual method further reduces the amount of computation required to achieve convergence.

Given that the negative regularized ML objective function that we minimize is convex, approximately solving the reduced unsymmetric primal-dual Newton equations is appropriate. Symmetrizing the unsymmetric system, while potentially useful for nonconvex problems, would in this case require solving for $2n$ variables without avoiding the potential for ill-conditioning. Our stabilization technique avoids the structural ill-conditioning of the condensed primal-dual matrix, and therefore solving the reduced system poses no asymptotic difficulty as the barrier parameter approaches zero. The computational efficiency and relative simplicity of formation of the reduced system of equations pose such a strong advantage that our choice of primal-dual method almost seems obvious for this problem.

Since Newton's method converges quadratically near the solution, for a well-conditioned system in the limit as $\mu \to 0$, one truncated-Newton step per subproblem should yield an increasingly accurate and well scaled direction to the subproblem solution for $\mu_k$. As $\mu$ is decreased, the subproblem solutions should become "close" to each other for a convex problem [13]. Yet, the example in Table 1 illustrates that the direction produced by the early-terminated CG can in fact become less accurate for smaller $\mu$ due to the structured ill-conditioning in $M$. In practice, we do not require the accuracy of the test example in Table 1. Our termination conditions are defined to be near the point on the trajectory where the stabilization approximation becomes accurate enough to guarantee descent. These termination criteria are quite accurate by the standards of the tomography community. Thus, although most reconstruction problems are unlikely to be *severely* affected by ill-conditioning, the potential for slow convergence near the solution due to ill-conditioning does exist. Our experience has been that stabilization has been an effective safeguard against poor performance for small values of the barrier parameter.

## 7 Acknowledgments

This study utilized the high-performance computational capabilities of the IBM RISC/6000 SP system at the Division of Computer Research and Technology, National Institutes of Health, Bethesda, MD. We are grateful to Jürgen Seidel of the Department of Nuclear Medicine, National Institutes of Health, for kindly providing us with the small animal data and Monte Carlo simulation data.

## References

[1] R.H. Chan and M.K. Ng, *Conjugate gradient methods for Toeplitz systems*, SIAM Review, 38 (1996), pp. 427–482.

[2] G. Chinn and S-C. Huang, *A general class of preconditioners for statistical iterative reconstruction of emission computed tomography*, IEEE Trans. Med. Imag., 16, (1997), pp. 1–10.

[3] N.H. Clinthorne, T-S. Pan, P-C. Chiao, W.L. Rogers, and J.A. Stamos, *Preconditioning methods for improved convergence rates in iterative reconstructions*, IEEE Trans. Med. Imag., 12 (1993), pp. 78–83.

[4] A.R. Conn, N. Gould, and Ph.L. Toint, *A primal-dual algorithm for minimizing a non-convex function subject to bound and linear equality constraints*, IBM T.J. Watson Research Center Technical Report 96/9, 1996.

[5] R.S. Dembo and T. Steihaug, *Truncated-Newton algorithms for large-scale unconstrainted optimization*, Math. Programming., 26 (1983), pp. 190–212.

[6] G. Demoment, *Image reconstruction and restoration: overview of common estimation structures and problems*," IEEE Trans. Acoust., Speech, Signal Processing, 37 (1989), pp. 2024–2036.

[7] A.P. Dempster, N.M. Laird, and D.B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm*, J.Roy .Stat. Soc., 39 (1977), pp. 1–38.

[8] A.R. DePierro, *A modified expectation maximization algorithm for penalized likelihood estimation in emission tomography*, IEEE Trans. Med. Imag., 14 (1995), pp. 132–137.

[9] ___*On the convergence of an EM-type algorithm for penalized likelihood estimation in emission tomography*, IEEE Trans. Med. Imag., 14 (1995), pp. 762–765.

[10] J. Dussault, *Numerical stability and efficiency of penalty algorithms*, SIAM J. Numer. Anal., 32 (1995), pp. 296–317.

[11] A.V. Fiacco and G.P. McCormick, *Nonlinear Programming: Sequential Unconstrained Minimization Techniques,* John Wiley and Sons, New York, 1968.

[12] A. Forsgren and P.E. Gill, *Primal-dual interior methods for nonconvex nonlinear programming*, Report NA 96-3, Department of Mathematics, University of California, San Diego, 1996.

[13] A. Forsgren, P.E. Gill, and J.R. Shinnerl, *Stability of symmetric ill-conditioned systems arising in interior methods for constrained optimization*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 187–211.

[14] G. Golub and C. Van Loan, *Matrix Computations*, second edition, Johns Hopkins University Press, Baltimore, 1989.

[15] P.J. Green, *On use of the EM algorithm for penalized likelihood estimation*, J.Roy. Stat. Soc. B., 52 (1990), pp. 443–452.

[16] ___, *Bayesian reconstructions from emission tomography data using a modified EM algorithm,* IEEE. Trans. Med. Imag., 9 (1990), pp. 84–93.

[17] C.C. Gonzaga, *Path-following methods for linear programming*, SIAM Review, 34 (1992), pp. 167–224.

[18] T. Hebert and R. Leahy, *A generalized EM algorithm for 3-D Bayesian reconstruction from Poisson data using Gibbs priors*, IEEE Trans. Med. Imag., 8 (1989), pp. 194-202.

[19] G.T. Herman, *Image Reconstruction from Projections: the Fundamentals of Computerized Tomography,* Academic Press, New York, 1980.

[20] M.R. Hestenes and E. Steifel, *Methods of conjugate gradients for solving linear systems*, J. Research of the National Bureau of Standards," 49 (1952), pp. 409–435.

[21] H.M. Hudson and R.S. Larkin, *Accelerated image reconstruction using ordered subsets of projection data*, IEEE Trans. Med. Imag., 13 (1994), pp. 601–609.

[22] F. Jarre and M.A. Saunders, *A practical interior-point method for convex programming*, SIAM J. Optim., 5 (1995), pp. 149–171.

[23] C.A. Johnson, *Nonlinear optimization for volume PET reconstructions*, Ph.D. dissertation, Department of Operations Research and Engineering, George Mason University, Fairfax, VA, 1997.

[24] C.A. Johnson, Y. Yan, R.E. Carson, R.L. Martino, and M.E. Daube-Witherspoon, *A system for the 3D reconstruction of retracted-septa PET data using the EM algorithm*, IEEE Trans. Nucl. Sci., 42 (1995), pp. 1223–1227.

[25] C.A. Johnson, J. Seidel, R.E. Carson, W.R. Gandler, A. Sofer, M.V. Green, and M.E. Daube-Witherspoon, *Evaluation of 3D reconstruction algorithms for a small animal PET camera*, *IEEE Trans Nucl. Sci.*, 44 (1997) pp. 1303–1308.

[26] L. Kaufman, *Implementing and accelerating the EM algorithm for positron emission tomography,* IEEE Trans. Med. Imag., 6 (1987), pp. 37–51.

[27] ___, *Solving emission tomography problems on vector machines*, Ann. Oper. Res., 22 (1990), pp. 325–353.

[28] M. Kojima, S. Mizuno, and A. Yoshise, *A primal-dual interior point algorithm for linear programming*, in Progress in Mathematical Programming: Interior-Point and Related Methods, N, Megiddo. ed., Springer Verlag, New York, 1989, pp. 29–47.

[29] A. Kuruc, *Probability measure estimation using "weak" loss functions in positron emission tomography*, in Tomography, Impedance Imaging and Integral Geometry, Lectures in Appl. Math. 30, M. Cheney, P. Kuchment, and E.T. Quinto, eds., American Mathematical Society, providence, RI, 1994, pp. 125–142.

[30] D.S. Lalush and B.M.W. Tsui, *The importance of preconditioners in fast Poisson-based iterative reconstruction algorithms for SPECT*, 1995 IEEE Nuclear Science Symposium Conference Record(1996), pp. 1326–1330.

[31] J. Liow and S. Strother, *Practical tradeoffs between noise, quantitation and number of iterations for maximum likelihood-based reconstructions*, IEEE Trans. Med. Imag., 13 (1991), pp. 601–609.

[32] K. Lange and R. Carson, *EM Reconstruction Algorithms for Emission and Transmission Tomography*, J. Comp. Assist. Tomogr., 8 (1984), pp. 306–316.

[33] K. Lange, M. Bahn, and R. Little, *A theoretical study of some maximum likelihood algorithms for emission and transmission tomography*, IEEE Trans. Med. Imag., 6 (1987), pp. 106–114.

[34] K. Lange, *Convergence of EM image reconstruction algorithms with Gibbs smoothing*, IEEE Trans. Med. Imag., 9 (1990), pp. 439–446.

[35] E. Levitan and G.T. Herman, *A maximum a posteriori probability expectation maximization algorithm for image reconstruction in emission tomography*, IEEE Trans. Med. Imag., 6 (1987), pp. 185–192.

[36] I. Lustig., R.E. Marsten and D.F. Shanno, *Interior point methods for linear programming: computational state of the art*, ORSA J. Comp, 6 (1994), pp. 1–14.

[37] G.P. McCormick, *The superlinear convergence of a nonlinear primal-dual algorithm*, Report T-550/91, Department of Operations Research, George Washington University, Washington, DC, 1991.

[38] E.U. Mumcuoglu, *Statistical Modeling and Fast Bayesian Reconstruction in Positron Tomography,* USC-SIPI Report 273, University of Southern California, 1994, dissertation.

[39] W. Murray, *Analytic expressions for the eigenvalues and eigenvectors of the Hessian matrices of barrier and penalty functions*, J. Optimization Theory App., 7 (1971), pp. 189–196.

[40] S.G. Nash, *Preconditioning of truncated-Newton methods*, SIAM J. Sci. Stat. Comput., 6 (1985), pp. 599–616.

[41] S.G. Nash and A. Sofer, *Block truncated-Newton methods for parallel optimization*, Math. Programming., 45 (1989), pp. 529–546.

[42] ___, *Assessing a search direction within a truncated-Newton method*, Oper. Res. Let., 9 (1990), pp. 219–221.

[43] ___,*A general-purpose parallel algorithm for unconstrained optimization*, SIAM J. Opt., 1 (1991), pp. 530–547.

[44] ___, *A barrier method for large-scale constrained optimizaton*, ORSA J. Comp., 5 (1993), pp. 40–53.

[45] ___, *Linear and Nonlinear Programming*, McGraw-Hill, New York, 1996.

[46] D.B. Ponceleon, *Barrier methods for large-scale quadratic programming*, Report SOL 91-2, Department of Operations Research, Stanford University, Stanford, CA, 1991, Ph.D. Thesis.

[47] L.A. Shepp and Y. Vardi, *Maximum Likelihood Reconstruction for Emission Tomography*, IEEE Trans Med. Imag., 1 (1982), pp. 113–122.

[48] E.M. Simantiraki and D.F Shanno, *An infeasible interior-point method for linear complementarity problems*, SIAM J. Optim., 7 (1997), pp. 620–640.

[49] D.L. Snyder, M.I. Miller, L.J. Thomas, Jr., and D.G. Politte, *Noise and edge artifacts in maximum-likelihood reconstructions for emission tomography*, IEEE Trans. Med. Imag., 6 (1987), pp. 228–238.

[50] Y. Vardi, L.A. Shepp, and L. Kaufman, *A statistical model for positron emission tomography*, J. Amer. Stat. Assoc., 80, (1985), pp. 8–37.

[51] D.W. Wilson and B.M.W. Tsui, *Noise properties of filtered-backprojection and ML-EM reconstructed emission tomography images*, IEEE Trans. Nucl. Sci., 40 (1993), pp. 1198–1203.

[52] M.H. Wright, *Interior methods for constrained optimization*, Acta Numerica (1991), pp. 341–407.

[53] ___, *Some properties of the Hessian of the logarithmic barrier function,* Math. Prog. 67 (1994), pp. 265–295.

[54] ___, *Ill-conditioning and computational error in interior methods for nonlinear programming*, Technical Report 97-4-04, Computing Sciences Research Center, Bell Laboratories, Murray Hill, NJ, 1997.

[55] S. Wright, *Stability of linear equation solvers in interior point methods*. SIAM J. Matrix Anal. App. 16 (1995), pp. 1287–1307.